Title: Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™

Author: M. Eduardoff T.E. Gross C. Santos M. de la puente D. Ballard C. Strobl C. Børsting N. Morling L. Fusco C. Hussing B. Egyed L. Souto J. Uacyisrael D. Syndercombe Court Á. Carracedo M.V. Lareu P.M Schneider W. Parson C. Phillips

PII: S1872-4973(16)30064-3
DOI: http://dx.doi.org/doi:10.1016/j.fsigen.2016.04.008
Reference: FSIGEN 1505

To appear in: *Forensic Science International: Genetics*

Received date: 8-12-2015
Revised date: 14-4-2016
Accepted date: 15-4-2016

Please cite this article as: {http://dx.doi.org/

Inter-laboratory evaluation of the EUROFORGEN Global ancestry-informative SNP panel by massively parallel sequencing using the Ion PGM™

M. Eduardoff[a,1], T.E. Gross[b,1], C. Santos[c], M. de la Puente[c], D. Ballard[d], C. Strobl[a], C. Børsting[e], N. Morling[e], L. Fusco[e], C. Hussing[e], B. Egyed[f], L. Souto[g], J. Uacyisrael[h], D. Syndercombe Court[d], Á. Carracedo[c,i], M.V. Lareu[c], P.M Schneider[b]; The EUROFORGEN-NoE Consortium; W. Parson[a,j], C. Phillips[c]*

[a] Institute of Legal Medicine, Medical University of Innsbruck, Innsbruck, Austria

[b] Institute of Legal Medicine, Faculty of Medicine, University of Cologne, Cologne, Germany

[c] Forensic Genetics Unit, Institute of Forensic Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain

[d] Faculty of Life Sciences and Medicine, King's College, London, UK

[e] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[f] Department of Genetics, Faculty of Science, Eötvös Loránd University Budapest, Hungary

[g] Department of Biology, University of Aveiro, Aveiro, Portugal

[h] Fiji Police Forensic Biology and DNA Laboratory, Nasova, Suva, Fiji

[i] Center of Excellence in Genomic Medicine Research, King Abdulaziz University, Jeddah, Saudi Arabia

[j] Forensic Science Program, The Pennsylvania State University, PA, USA.

[1] These authors contributed equally to the work.

* Corresponding author. Tel.: +34 981 583 015.

E-mail address: c.phillips@mac.com (C. Phillips).

Evaluation of the Global AIM-SNP panel with Ion PGM™

Global AIMs markers

↓

97.6% assay conversion rate

↓

Inter-laboratory validation (**5** collaborators)

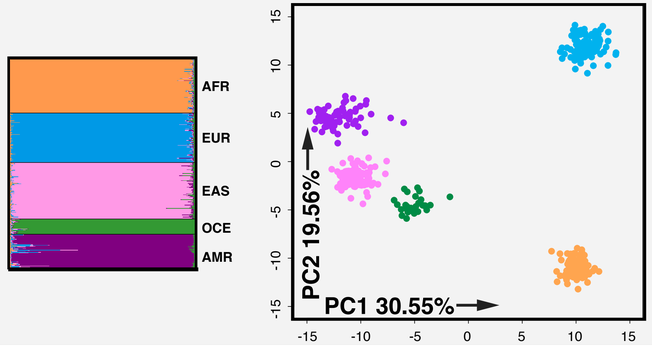
↓

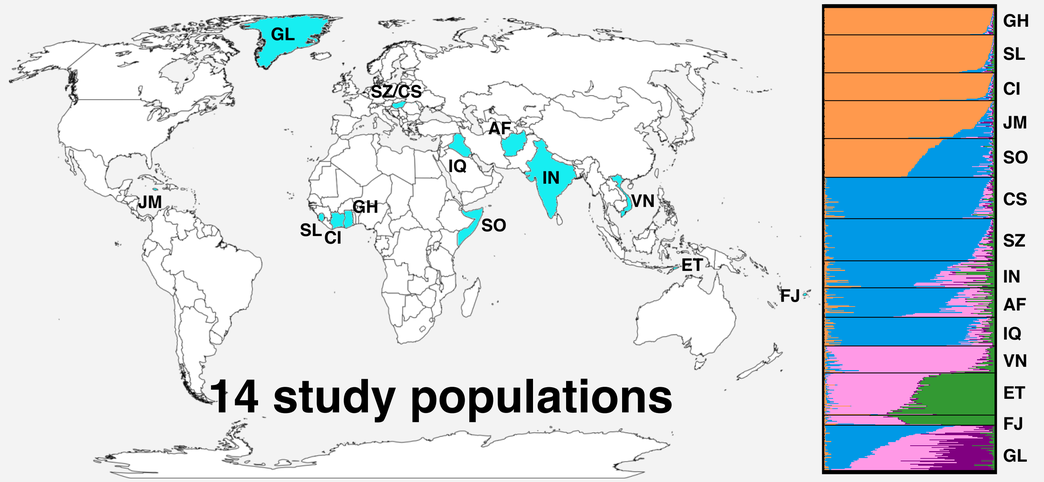
inter-laboratory database | concordance **> 99.8%**

full concordance down to **100pg of DNA**

mixtures detected up to **1:9** ratio

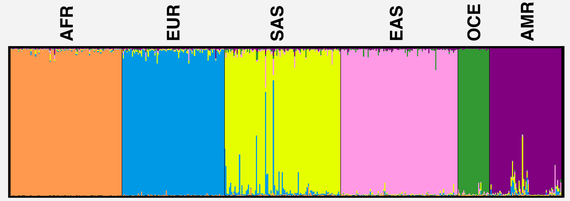
Designed for **5 population groups**

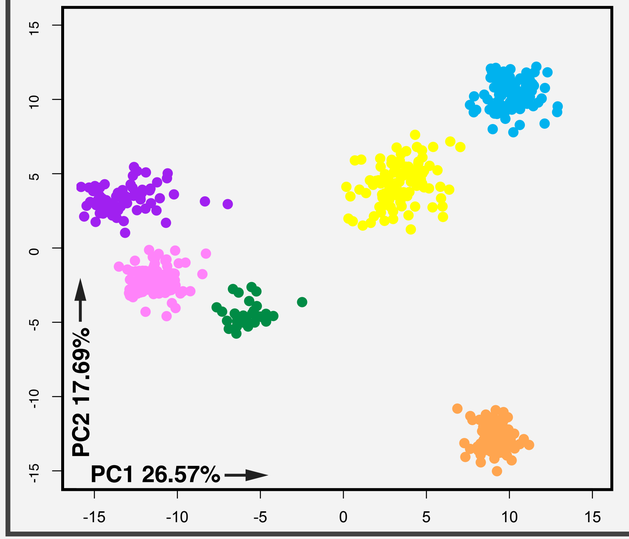




14 study populations

Possibility to differentiate up to **6 population groups**





Highlights

● The first custom-built forensic MPS multiplex was built for the EUROFORGEN Global ancestry-informative SNP panel analyzed with the Ion PGM™ system.

● 125 of 128 originally selected SNPs were successfully incorporated: a 97.6% assay conversion rate. Three substitute SNPs were added, but one SNP failed to provide usable sequence reads.

● Five-laboratory evaluations of the assay assessed sequencing performance, forensic sensitivity, and mixture detection.

● Studies of 14 novel populations indicate good informativeness for five continental population group differentiations and admixed populations, although distinguishing South Asian populations still requires extended ancestry-informative SNP panels.

Abstract

The EUROFORGEN Global ancestry-informative SNP (AIM-SNPs) panel is a forensic multiplex of 128 markers designed to differentiate an individual's ancestry from amongst the five continental population groups of Africa, Europe, East Asia, Native America, and Oceania. A custom multiplex of AmpliSeq™ PCR primers was designed for the Global AIM-SNPs to perform massively parallel sequencing using the Ion PGM™ system. This study assessed individual SNP genotyping precision using the Ion PGM™, the forensic sensitivity of the multiplex using dilution series, degraded DNA plus simple mixtures, and the ancestry differentiation power of the final panel design, which required substitution of three original ancestry-informative SNPs with alternatives. Fourteen populations that had not been previously analyzed were genotyped using the custom multiplex and these studies allowed assessment of genotyping performance by comparison of data across five laboratories. Results indicate a low level of genotyping error can still occur from sequence misalignment caused by homopolymeric tracts close to the target SNP, despite careful scrutiny of

candidate SNPs at the design stage. Such sequence misalignment required the exclusion of component SNP rs2080161 from the Global AIM-SNPs panel. However, the overall genotyping precision and sensitivity of this custom multiplex indicates the Ion PGM™ assay for the Global AIM-SNPs is highly suitable for forensic ancestry analysis with massively parallel sequencing.

Keywords: Massively parallel sequencing (MPS); Ion PGM™; Ancestry-informative SNPs; Forensic ancestry analysis

1. Introduction

Developing assays of ancestry informative markers (AIMs) is of particular interest in forensic genetics as they can provide investigative leads in cases where the source individual is not known. Studies using many hundreds of markers suggest worldwide populations can be placed in groups based on genetic similarity, closely corresponding to their continental distribution [1-4]. However, this pattern can be highly dependent on the sampling scheme, as much worldwide genetic diversity takes the form of geographic clines [5-7]. For forensic ancestry analysis, differentiation of five population groups, comparing Africa, Europe, East Asia, Native America, and Oceania, is a practical objective using small-scale marker sets selected to have strongly contrasting allele frequencies [8]. Two objectives have been proposed for forensic ancestry inference and the estimation of co-ancestry proportions in admixed individuals: i. assembling small marker sets targeting the highest possible allele frequency divergence values during SNP selection [9,10]; and ii. balanced divergence amongst the target population groups to differentiate each with equal power as a way to reduce estimation bias of co-ancestry proportions in admixed individuals [8].

Massively parallel sequencing (MPS) has the capacity to greatly enhance forensic DNA analysis by providing accurate sequence data for hundreds of loci resulting in a marked increase in the information gained from a single DNA test [11-15]. Initial assessments of MPS indicate that sequence data with sufficiently high coverage and

reliable genotypes can be produced for most loci. However, careful scrutiny of the sequence characteristics is required for each SNP chosen. Furthermore, the sequence data analysis systems of MPS platforms developed for forensic use are still not fully developed and their further optimization is necessary before any MPS multiplex can be introduced in the forensic field [11,13,16].

The EUROFORGEN Global ancestry-informative single nucleotide polymorphism panel (herein Global AIM-SNPs) comprises 128 markers designed to distinguish the five continental groups outlined above [17]. The Global AIM-SNPs were compiled to provide the key characteristic of a balanced differentiation of each population group, i.e. the cumulative SNP variation has equal levels of population divergence amongst the five groups so that admixture proportions, detected as co-ancestry in the individual, are estimated with minimum bias. In the evaluation study reported here, we have assessed a custom primer set for the Global AIM-SNPs developed for the Thermo Fisher Scientific (TFS) Ion Personal Genome Machine® (PGM™) system [18]. The proportion of Global AIM-SNPs incorporated into the custom-made AmpliSeq™ primer set gives a first indication of the assay conversion rate (i.e. the number of user-selected SNPs successfully incorporated) that can be expected for the Ion PGM™ system in forensic use. The assay conversion rate for MPS systems is important to assess with regard to the much larger PCR multiplexing levels possible with this technology. It is also necessary to gauge how easily novel SNP discoveries for such purposes as forensic phenotyping [19] or specialized ancestry analyses [20,21] can be incorporated into single multiplexes for MPS analysis. Furthermore, there are initial indications that well optimized SNaPshot-based forensic SNP PCR multiplexes can be easily combined and ported directly to MPS with little or no modification [12].

Once the Global AIM-SNPs PCR multiplex had been successfully prepared by TFS, evaluations were made between five EUROFORGEN laboratories (affiliations a-e). Evaluations considered: i. component SNP performance in MPS; genotyping precision and concordance; ii. gauging the assay's forensic sensitivity by analyzing simple dilution series and degraded DNA plus detection of mixed DNA; and iii. the ancestry differentiation power of those Global AIM-SNPs successfully incorporated into the assay and giving reliable sequencing data. Assessments of the assay's

forensic performance followed a previously established framework used to evaluate the HID-Ion AmpliSeq™ Identity Panel [13]. This simple framework consisted of genotyping DNA dilutions, degraded DNA, artificial mixtures and universal control DNAs having publicly-available genotypes for the same SNPs generated from alternative MPS SNP genotyping techniques.

In common with the principal findings of the previous evaluation of the HID-Ion AmpliSeq™ Identity Panel [13], our results indicate the biggest obstacle to successful MPS genotyping of forensic SNPs is the problem of misalignment of detected sequences due to closely sited homopolymeric tracts or Indels with consequent allele miscalling. Therefore, in addition to selecting SNPs with optimum properties for a particular forensic test, in this case ancestry analyses [17], very careful scrutiny of flanking sequence characteristics is required to avoid impairing the test's genotyping precision using MPS technology.

2. Material and methods

The term *run* is used for a combination of multiple samples on a single Ion PGM™ sequencing chip. The term *analysis* refers to the sequencing results of a specific sample from one run. Within the Ion PGM™ analysis software the term *allele frequency* is used to describe the sequencing read counts for each allele per marker as previously described [13], but is easily mistaken for the population genetics term, therefore we opted to use *allele read frequency* (ARF) in the following experimental descriptions. In the following description of population analyses the metric Divergence is capitalized to distinguish it from the phenomenon of population divergence.

2.1. DNA samples and population data

For genotyping concordance studies, seven Coriell control DNAs were selected with an origin from one of the five population groups differentiated by the Global AIM-SNPs. Coriell control DNA samples comprised the CEPH Utah European trio of

NA06994, NA07000; and NA07029; Yoruba African NA18498; Han-Chinese HG00403; Melanesian NA10540; and Quechuan from Peru NA11200. Use of Coriell control DNA samples enabled comparison of three independent MPS SNP genotyping techniques, as their genotypes are in the online databases of 1000 Genomes (using Illumina HiSeq [22]) and Complete Genomics (using an in-house DNA nanoarray method [23]). For reference purposes the standard forensic control DNA 9947A, available to most forensic laboratories, acted as a universal control.

Genotypes of component Global AIM-SNPs were obtained from three sources: i. 1000 Genomes Phase III data [24]; ii. the Stanford University HGDP-CEPH analyses [3], accessed with *SPSmart* [25]; and iii. genotypes generated in this study for in-house population samples of interest. All population descriptions are detailed in Table 1.

We opted to use single reference populations from 1000 Genomes without high levels of admixture and with low intra-population variability [26], consisting of AFR (ESN), EUR (GBR) and EAS (JPT) groups plus sets of two CEPH OCE populations and five CEPH AMR populations (Table 1, population no. 1-6). The 1000 Genomes GIH population was chosen as the reference data for analysis of South Asia region test/study populations. ESN, GBR, JPT and GIH populations gave the lowest within population average pairwise genotype differences (Fig. 4A, orange cells) from population analyses described in section 3.6. This strategy also compensated for the large contrasts in sample size of 1000 Genomes data for the first three groups alongside the much smaller sample sizes of Oceanian and Native American populations, which can interfere with *STRUCTURE* analyses [26].

Other unadmixed populations from 1000 Genomes were used as test sets (Table 1, population no. 7-28). South Asian populations (Table 1, population no. 6 and 19-22) were included in certain analyses to assess the Global AIM-SNPs' capacity to differentiate Europeans and South Asians, which are less divergent than the five population groups the Global AIMs panel was designed to differentiate.

Study populations comprised 551 samples of 14 different in-house populations (Table 1, population no. 29-42) selected to broaden the geographic scope of population data

already compiled for the Global AIM-SNPs. All study population samples were obtained with informed written consent for scientific research. Institutional ethical approval documents were obtained for all study population samples, and subjected to a subsequent review by the European Commission. DNA was extracted with the QIAamp® DNA Mini Kit, EZ1 DNA Investigator Kit (both Qiagen) or as described in [27].

2.2. Preparation of DNA for MPS

The Ion AmpliSeq™ Library Kit 2.0 was used to prepare DNA libraries, following manufacturer's guidelines [28]. Apart from the forensic sensitivity evaluations (dilution series, degraded DNA samples and mixtures), all libraries were prepared from 1-10 ng DNA input. While the majority of samples were amplified with full volumes in all reaction steps (herein the full volume protocol) as recommended by manufacturer's guidelines [28], another portion was amplified with half volumes in all reaction steps of library preparation (the half volume protocol). Irrespective of reaction volumes used, samples were subject to 18 or 21 target amplification cycles for 10 or <10 ng DNA input, respectively. DNA libraries were quantified with either Ion Library TaqMan® Quantitation Kit or Qubit® ds DNA HS Assay Kit following manufacturer's guidelines [28]. Use of the term '+5' denotes the 5-cycle DNA library re-amplification applied prior to Qubit® quantification in the Ion AmpliSeq™ Library Kit 2.0 protocol [28]. Ion Library TaqMan® quantitation does not require this amplification step. In addition, libraries of the control DNA sample set and 82 population samples were prepared with full and half volume protocols to compare both preparations.

For barcoding and purification of all DNA libraries, Ion Xpress™ Barcode Adapters and Agencourt AMPure XP magnetic beads were used following manufacturer's guidelines [28]. Libraries were pooled for a final concentration of 8-12 pM for template preparation with the Ion OneTouch™ 200 Template Kit v2 and the success of this step was evaluated with Ion Sphere™ Quality Control Kit following manufacturer's guidelines [29]. Sequencing made use of Ion PGM™ Sequencing 200 Kit v2 and Ion 316™ or 318™ v2 chips according to manufacturer's guidelines [30].

The sensitivity of the Global AIMs panel was assessed with a dilution series of Coriell control DNA NA07000 with 10 ng (18 cycles), 1 ng (21 cycles), 500 pg (21 and 25 cycles), 250 pg (21 and 25 cycles), 100 pg (21 and 25 cycles), 50 pg (25 and 25+5 cycles), 25 pg (25 and 25+5 cycles), and 10 pg (25+5 cycles) DNA input. Two challenging DNA samples were also analyzed: Bone 1 and Bone 2, both previously assessed for inhibition/degradation. Quantifiler Duo tests suggested Bone 1 was inhibited, as the internal PCR control (IPC) required more than 31 cycles. Quantifiler Trio tests gave small target concentration/large target concentration ratios of Bone 1=2.45 and Bone 2=5.49, suggesting degraded DNA in both cases, but no IPC indications of inhibition for either sample. The two bone samples were prepared with 21+5 cycles.

The different numbers of PCR cycles for varying DNA inputs as well as the additional 5 amplification cycles after library preparation for DNA libraries with low yields (below 100 pM) were chosen following manufacturer's guidelines for MPS analysis of limited samples [28].

To evaluate DNA mixtures, two Coriell DNA samples of known origin, NA07000 (SS1, European, 8.30 ng/µl) and NA184984 (SS2, Yoruba African, 7.76 ng/µl), were quantified using Qubit® ds DNA HS Assay Kit and mixed in volume ratios of 1:9; 1:3; 1:1; 3:1 and 9:1. Each mixture DNA library was sequenced twice in different runs (replicates A and B).

2.3. Data analysis

Raw sequencing data files were processed with Torrent Suite™ 4.2 (herein TS) and HID_SNP_Genotyper plugin version 4.2 (herein Genotyper) with germline low stringency parameter settings [31]. A target and a hotspot bed file (Supplementary Files S1 and S2) were used for identification of Global AIM-SNPs with genome build hg19 (GRCh37). Both Genotyper report files (csv and vcf of each sample) were used for further analysis and data was processed downstream with R [32] (v. 3.0.3) or Excel.

2.4. Criteria for marker or sample data exclusion and manual correction of genotypes

During the concordance analysis of the control DNA sample set and the preliminary check of population sample genotypes problems were observed for four Global AIM-SNPs and some population samples, e.g. population specific variants, that appeared to cause a high number of no-call (NN) genotypes or alignment difficulties due to closely sited homopolymeric tracts. A thorough analysis of corresponding raw sequencing output (BAM files in IGV [33,34]), in combination with appropriate vcf files resulted in measures introduced for manual correction of genotypes for SNPs rs595961, rs6875659 and rs12402499, but the exclusion of rs2080161. More details on manual correction and exclusion of these four SNPs are provided in section 3.2.4, Supplementary Table S1 and Supplementary File S3.

Genotypes of population samples were reviewed prior to analysis of data to prevent bias from the collection of data from underperforming SNPs or samples. Underperforming SNPs were defined as having a higher than average no-call rate due to low quality values or low sequence coverage (i.e. more than the average 1.2 no-call genotypes per SNP per 551 samples, 0.21%). Some no-call genotypes were occasionally observed in good quality samples, but higher numbers of no-call genotypes tended to indicate problems with sample DNA quality. Therefore, population samples with less than 95% complete genotypes (122/128) were excluded from any further analysis (data not shown); a higher stringency than the 90% complete genotypes threshold used in a similar study [35]. In the majority of these underperforming samples a high number of no-call genotypes also involved low average sequence coverage per sample. Consequently, vcf files of population samples with an average coverage per sample below 200x, but less than 5% no-call genotypes were scrutinized to ensure the genotypes available for population analysis were reliable. Furthermore, any genotype calls with sequence coverage less than 30x were either confirmed or rejected by review of the sample's vcf data. Sequence coverage thresholds comprised a minimum coverage (total reads) of 20x for heterozygotes and 10x for homozygotes with a minimum coverage per allele and strand direction

(number of forward or reverse reads per allele) of 10x or 5x, respectively. If coverage was less or reads per allele were not within allelic balance (40-60% for heterozygotes, 90% for homozygotes) or strand bias thresholds (25-75%), as previously established [13], then genotypes were manually corrected to no-calls. For example, a homozygote call with 15x coverage comprising 3x forward strand coverage and 12x reverse strand coverage was manually corrected to a no-call, whereas a homozygote call with 15x coverage split into 7x forward coverage and 8x reverse coverage was maintained.

2.5. Population analyses

Shannon's Divergence values were calculated for each SNP using the cross-validation option in *Snipper* [36] from pairwise and one-against-all population comparisons. Divergence values were converted to Rosenberg's informativeness-for-assignment metric I_n [9] ($\log_n(2)$ values from multiplication by 0.69) and the final population specific Divergence (PSD) for each population group was obtained from their cumulative Divergence values. The *Snipper* portal was also used to calculate classification likelihoods ratios (LR) by uploading an Excel file of reference data.

Population analyses with *STRUCTURE* v. 2.3.4 [37] were performed following previous guidelines [38]. One to nine populations (K=1 to K=9) were assumed and five replicate analyses were executed for each K value. The analyses were performed considering the admixture ancestry model with correlated allele frequencies. Each analysis run consisted of 100,000 burnin steps followed by 100,000 MCMC steps to achieve accurate estimation of posterior probabilities. The optimum K value was estimated by computing results with Structure Harvester [39] and following previous guidelines [40]. *STRUCTURE* ancestry membership proportions were plotted using a combination of CLUMPP v. 1.1.2 [41] and distruct v. 1.1 [42]. PCA analyses were performed using R software v.3.1.2 [32] and executing a homemade script (available on request).

Population allele frequencies, average number of pairwise genotype differences within or between populations, F_{ST} calculations and exact tests of Hardy-Weinberg equilibrium (HWE) were performed using Arlequin v. 3.5 [43].

3. Results and discussion

3.1. Ion PGM™ custom assay design and conversion rate

To the best of our knowledge, the EUROFORGEN Global AIM-SNPs panel is the first *custom* forensic multiplex design (i.e. an end-user's own marker selection) compiled by TFS with AmpliSeq™ primers for forensic SNP analysis with the Ion PGM™. It is important to stress that manufacturers of MPS systems have developed forensic SNP tests from established panels, but these have gone through a series of optimization steps requiring major adjustments of component marker combinations (ME, TEG, WP, CP; personal communication with manufacturers). Therefore, the assay conversion rate achieved for the original selection of 128 SNPs is an important indicator of how readily future SNP selections could be adapted into the preparatory target amplification ahead of MPS analysis. Although careful scrutiny was made of context sequence for the original Global AIM-SNP candidates, three SNPs: rs5757362, rs2282107 and rs7246968 presented insurmountable problems for primer design. All were positioned in repeat regions that would create non-specific primer binding and significant amounts of off-target sequence reads. Additionally, rs2282107 and rs7246968 were too close to long homopolymeric tracts preventing efficient sequencing. Two of three substitute SNPs: rs2837352 (for rs5757362) and rs16946159 (for rs2282107) had similar population differentiation properties but were sited in different regions, so were successfully incorporated. The remaining substitute SNP of rs7250345 (for rs7246968) was in the same region so had near-identical allele frequencies. However, the closely sited repetitive sequences comprised very long SINE elements so this whole region was abandoned and substitute rs11048128 was successfully incorporated instead.

The changes in individual and cumulative population specific Divergence (PSD) values for the above Global AIM-SNP substitutions (and excluding underperforming rs2080161) are detailed in Supplementary Table S1. From just three replacements amongst 128 SNPs, changes are marginal, although there is a noticeable drop in the

cumulative PSD value for the East Asian group. Overall, we were satisfied with the Ion PGM™ assay conversion rate of 97.6%, although as described in section 3.2.4, rs595961, rs6875659 and rs2080161 all produced alignment problems due to closely sited homopolymeric tracts, with SNP rs2080161 excluded from the panel because of unreliable data. We would have expected such flanking sequence problems to be largely identified during the TFS primer and multiplex design process in each SNP's context sequence review.

3.2. Genotyping concordance

Assessment of genotype concordance was made in three ways: i. comparing genotypes from identical control DNA samples prepared and run in different laboratories (inter-lab concordance, 37 analyses); ii. comparing Ion PGM™ genotypes to those in 1000 Genomes and Complete Genomics public databases for Coriell control DNA samples; and iii. comparing runs of the same sample with full and half volume protocols (8 control DNA sample analyses and 82 populations sample analyses). To allow for varying numbers of replicates for different samples and varying numbers of no-calls, the concordance rates for individual samples are based on the number of called genotypes.

3.2.1. Inter-lab concordance

Inter-lab concordance of called genotypes was 99.81% (4707/4716), with a no-call rate of 0.42% (20/4736). Discordances were observed in SNPs rs6875659, rs2080161, rs9934011 and rs9908046 in three different samples, as described in Table 2, producing a discordance rate of 0.19% (9/4716).

3.2.2. Concordance between Ion PGM™ genotypes and online databases

Genotypes of all 128 Global AIM-SNPs are listed by 1000 Genomes, but for only four Coriell control DNA samples (NA06994, NA07000, HG00403, NA18498, 19 analyses). Comparisons with 1000 Genomes genotypes resulted in a no-call rate of 0.58% (14/2432) caused by 8 different SNPs (see section 3.5.3 and Supplementary

Table S1, rows 6, 9-11, 16-18, 22, column K) and a concordance rate of 99.84% (2414/2418). Four discordant genotypes (0.16% discordance) were observed in four different analyses of the same Coriell control sample, all in SNP rs6875659, as shown in Table 2.

Five Coriell control DNA samples (the above plus NA07029) are listed by Complete Genomics; so concordance rates are based on 2944 genotypes from 23 analyses. The no-call rate was 0.47% (14/2944); while 99.86% of called genotypes (2926/2930) were concordant. The same SNP causing the discordances with 1000 Genomes data resulted in 0.14% discordance rate with Complete Genomics data (see Table 2).

3.2.3. Concordance between full and half volume protocols

The comparison of full and half volume library preparation protocol to analyze the control DNA sample set gave a high concordance rate of 99.95% (1/2038 discordant genotype in rs2080161) and a no-call rate of 0.49% (10/2048).

In addition, genotypes between protocols were compared in 82 samples from seven populations. Of these, 41 had genotype differences (no-calls and discordant genotypes) comparing full and half volume protocols. Thirty-two analyses (76.19%) had up to three differences in SNPs rs595961 and rs2080161 (see section 3.5.2) or in SNPs with high no-call rates (>4 no-calls in 551 genotypes): rs4979274, rs499827, rs310644 and rs1366220 (see section 3.5.3). Among these six SNPs, only rs595961 produced discordant genotypes (11), with others giving no-call genotypes in one of the analyses. Overall, of 31 no-call genotype differences observed, 14 (45%) were in full volume protocols. In another 7 samples, more than five genotype differences were observed, mainly no-calls and from low sequence coverage in the half volume protocol. One African sample produced a discordant genotype in rs2814778 using the half volume protocol, showing 11% of T bases in forward reads in contrast to 100% C bases on both strands in the full volume protocol. Visual scrutiny of sequence output (BAM file data) suggested an emulsion PCR incorporation error or misalignment. In summary, the full volume protocol gave more useable SNP data in 14 samples, while the half volume protocol was better in 11 (44%). Therefore, typing of samples with

the half volume protocol is a feasible strategy that reduces sequencing costs with very minor or no loss of data quality.

3.2.4. Manual correction of genotypes

Concordance analyses indicated genotyping problems for four SNPs. These problems were further investigated to find appropriate measures for manual correction of genotypes or exclusion of a component SNP, if reliable genotype calls for that marker could not be guaranteed. Genotype calls for rs2080161 , rs595961 and rs6875659 were affected by homopolymeric tracts close to the target SNP position (Supplementary Table S1 row 4-6, column U and Supplementary File S3, SNPs 1-3), causing either misalignment or truncation of reads before they reached the SNP position. In rs2080161, several poly-T tracts are found within the amplicon in both directions, resulting in unreliable genotype calls. Therefore, component SNP rs2080161 was excluded from the panel.

In contrast, rs595961 and rs6875659 had severely affected allele calls in one strand direction, while the opposite direction had well balanced sequence coverage of each allele and reliable genotype calls (assessed by visual scrutiny of individual BAM files, Supplementary File S3, SNPs 2 and 3). Therefore, genotypes for rs595961 and rs6875659 were corrected manually by inferring allele calls and genotypes only from forward or reverse reads respectively, based on read counts in their vcf files. It is notable that for rs6875659, African samples (common allele A) were more prone to incorrect allele calls compared to the European, East Asian, South Asian or Oceanian samples we genotyped (major allele G) [44]. Any A nucleotide genotype call creates a run of five successive As with the rs6875659 allele being the last A in the forward direction, causing discordances in 4 of 5 analyses of Coriell control DNA NA18498 in this SNP (see Table 2). The applicability of genotype correction measures was verified for rs595961 and rs6875659 by comparing corrected Ion PGM™ genotypes from Coriell control samples to 1000 Genomes and Complete Genomics data. Similarly, comparison of Somali population genotypes from Ion PGM™ to those obtained independently using Sequenom® highlighted the effect of strand-specific misalignment on the reliability of genotype calls from MPS.

It is important to emphasize that manual correction of individual genotype calls is neither straightforward nor desirable, but continuous improvement of MPS analysis software is likely to address many of the observed issues, e.g. deducing genotype calls from only one strand direction directly within the appropriate analysis module.

Although no discordances were observed for rs12402499, a very high number of no-call genotypes in African samples (22%, 44/204) characterized the data from this SNP. Scrutiny of the context sequence of rs12402499 revealed a population-specific Indel (rs146348214, TTGA/-, Chr 1:101528955-101528958,) directly adjacent to the SNP position with ~15% frequency in Africans [45]. Both variants were sequenced without problems, but were incorrectly identified as a single variant within Genotyper, resulting in a no-call for rs12402499 in affected samples (Supplementary File S3, SNP 4). The vcf files of these samples were manually revised and no-call genotypes corrected for population analysis samples to compensate for this software error.

3.3. Sensitivity of Global AIM-SNPs assay and analysis of degraded DNA

Full concordance was observed in the NA07000 Coriell control sample dilutions in the DNA input ranges 10 ng to 100 pg using both 21 and 25 PCR cycles. SNP rs715605 was the only exception, with no-calls recorded from 100 pg input or less due to low coverage. SNP rs187153 gave no-call genotypes with 50 pg input or less. As expected, no-call genotypes, allele drop-ins and allele drop-outs, as well as locus drop-outs all rose in frequency with decreasing DNA inputs, but only below 100 pg (Table 3). It is interesting to note that samples amplified with an additional 5 PCR cycles after library preparation (+5) did not show increased sensitivity but actually had higher numbers of incomplete or missing genotypes. However, the one sample typed with 10 pg input DNA still produced 48% (62/128) concordant SNP genotypes.

DNA sample Bone 1 gave no genotypes for the analyzed markers. Bone 2 (input DNA=726 pg) produced four no-calls but gave average sequence coverage of 430x. This number of no-call genotypes is higher than the no-call rates seen in the dilution series samples with a similar input amount.

3.4. Mixtures

AIM-SNPs are usually selected to have highly skewed allele frequencies between populations or can even approach fixation in some populations (i.e. allele frequencies close to 0 or 1). Therefore, they have an impaired ability to detect mixtures of individuals with shared ancestry compared to most identity-informative SNPs, which have minor contrasts in allele frequencies across populations. However, mixtures of individuals with different ancestries will tend to show higher heterozygosities (% heterozygous loci in the profile) in AIM-SNPs. The comparison of heterozygosity levels of the two single-source DNA samples SS1 and SS2 to the expected DNA mixture produces a 40% increase in heterozygosity (Fig. 1A). The expected DNA mixture profile was based on the combination of both single-source sample genotypes. Comparison of the different mixture ratios to the expected DNA mixture profile shows that, while the 1:1 ratio is close to the expected heterozygosity value, there is a decrease in heterozygosity moving towards the most asymmetrical ratios due to non-detection (drop-out) of the minor allele (Fig. 1B).

The asymmetric distribution of no-call genotypes and minor allele drop-outs in different mixture ratios (1:9 vs 9:1 and 1:3 vs 3:1) suggests a slightly higher concentration of the first European NA07000 mixture component (SS1). Detailed concordance analysis of allele drop-outs between replicates indicate drop-out was not due to a PCR loss of an allele, but because the minor allele did not reach the 0.1 threshold of the *minimum_allele_frequency* parameter. Therefore, data was re-analyzed with *minimum_allele_frequency* adjusted to 0.02, as previously tested [13]. Results in Fig. 1C show that Genotyper detects more minor alleles with these adjusted settings. In fact, 90% of the allele drop-outs occurring in all ratios with default parameter settings were detected applying the 0.02 threshold. It is also noticeable that no-call genotypes in mixture analysis were mainly due to a small number of underperforming SNPs, comprising: rs4979274, rs310644, rs11048128 and rs7151991 (see Supplementary Table S1, row 10-11, 16-17, column O and section 3.5.3). In addition, rs12402499 gave no-call genotypes in all mixture ratios with the

African NA18498 (SS2) as major component, due to the population specific deletion described above (see section 3.2.4 and Supplementary File S3, SNP 4).

Fig. 2 shows deviation from expected ARF values (observed minus expected ARF) for alleles present in the European SS1 sample in different mixture ratios for 123 SNPs (excluded SNP rs2080161 and four triallelic loci with three alleles in the expected mixture genotypes were not used). Overall, the correlation between expected and observed ARFs is higher than 95% in all cases, (R^2 values shown in boxes below the plot). However, there is a discernable trend of higher deviations from expected ARFs in the more balanced mixture ratios.

Graphical representation of allele read frequencies for each SNP was previously shown to be a useful aid to identification of mixtures, as most frequencies will be displaced from their typical single-source patterns. The mixture ARFs are plotted in Fig. 3 with the single-source components SS1 top left and SS2 bottom right. The mixed ARF plots in Fig. 3 show the displacement effect is more pronounced in mixtures closer to a balanced 1:1 ratio, with replicate runs showing well matched patterns. It is possible to track the frequency of the European-specific alleles in SS1 that decrease with the mixture ratio, reaching a minimum of approximately 0.05% in the 1:9 ratio.

Given the expected and observed increase of heterozygosity levels in mixed DNA, it is important to explore the extent to which individuals with co-ancestry (from population admixture) can be differentiated from the SNP data of mixed-source DNA. Individuals with co-ancestry will show an equal degree of raised heterozygosity. The observed levels of average heterozygosity in 1000 Genomes unadmixed Africans and Europeans compared to those of admixed population samples are shown in Supplementary Table S3. This data shows individuals with co-ancestry have raised heterozygosity some 20-80% higher than individuals with single ancestry. However, because of the high correlation between input DNA and observed ARF values found in mixtures, it is relatively straightforward to distinguish them from mixed DNA samples: ARFs of admixed individuals will show patterns like those of single-source samples (see SS1 and SS2 in Fig. 3). In addition, information that a forensic sample is a mixture will be obtained primarily from STR typing routinely applied to all

casework. Proceeding analyses can then be made to de-convolute the SNP data to identify the ancestry of the contributors.

The six triallelic SNPs in the Global AIMs panel provide an additional way to identify mixtures. However, the detection of three alleles in triallelic SNPs offered by MPS is not accommodated in the automatic calling of Genotyper. Knowing that the lowest expected ARF of the minor allele in triallelic SNP genotypes will match those of biallelic SNPs (expected ARFs outlined in Supplementary Table S4), the observed ARF values in mixtures are likely to be similar. However, at more extreme mixture ratios, the expected ARF of the minor allele will be much lower, so care is required to avoid confusing such alleles with misincorporated nucleotides. In four of the six triallelic SNPs, a three-allele genotype was expected (rs2184030, rs4540055, rs433342 and rs17287498). We could reliably detect the minor allele in the mixture, for all ratios, by scrutiny of the accompanying ARFs. Even in the case of the 9:1 mixture, it was possible to detect the minor allele with an ARF of 3%; a reasonable match to the expected value of 5%.

3.5. Overall evaluation of component SNP performance

Identification of underperforming SNPs was based on the results of concordance, sensitivity and mixtures analyses of control DNAs plus the review of population sample genotypes, as well as evaluation of the key parameters: sequence coverage; allele read frequency; nucleotide misincorporation rates; and strand bias per allele. SNPs were assigned to one of the following three categories; i. SNPs with discordant genotypes; ii. SNPs with no-call genotypes; and iii. SNPs with good performance and high genotyping reliability. More details on this SNP categorization applied to all 128 Global AIM-SNPs are summarized in Supplementary Table S1.

3.5.1. Key sequence quality parameters

Evaluation of SNP performance based on the above four key MPS parameters was made from values averaged over all population samples.

The most important parameter and a key limiting factor of MPS analyses is sequence coverage. Within the population sample sets a minimum value of 106x and a maximum of 1647x sequence coverage were obtained. This substantial variation in coverage across the Global AIM-SNPs is likely due to differing PCR amplification efficiencies within the 128-plex PCR and has been previously described for Ion PGM™ SNP panels of similar size [11,13,15]. Nevertheless, 95% of Global AIM-SNPs (122/128) showed an average sequence coverage of more than 300x. A higher number of no-call or discordant genotypes (>1% no-call genotypes in 551 samples) were observed for the remaining six SNPs along with a lower average coverage (see Supplementary Table S1, row 6, 9-12 and 14, column P and T); matching results found in the genotyping concordance analyses (see section 3.2).

Another key factor in forensic SNP analysis is allelic balance, critical for reliable genotyping of heterozygotes as well as identifying mixed-source samples. The ARF parameter in MPS equates to signal ratios in heterozygotes detected by capillary electrophoresis. Apart from five markers, 123 Global AIM-SNPs gave ARF value ranges well within previously established thresholds [13] of >90% for homozygotes and 40-60% for heterozygotes. Four of those five SNPs showed mean ARFs only slightly higher than the threshold cutoff (61-65%). A marked deviation was observed for the single outlier SNP rs310644 (Supplementary Table S1, row 11, column Q). SNP rs310644 was also identified as a low coverage marker and had a high number of no-call genotypes in both concordance and mixture studies.

In addition to allelic balance, the nucleotide misincorporation rate, describing the percentage of non-allelic nucleotide calls from all sequence reads, is a key factor in the reliable genotyping of SNPs and identification of minor alleles in mixed DNA samples. The misincorporation rate was less than 1% for all but two SNPs: rs595961 with 2.9% misincorporation and rs2789823 with 1.8%. However, examination of context sequence data in both SNPs indicated the apparent nucleotide misincorporation was actually caused by a small proportion of misaligned reads due to homopolymeric tracts close to the target sites (Supplementary Table S1, row 4 and 68, column S).

Lastly, strand bias per allele, measuring the ratio of sequence reads of one allele for each strand direction, can significantly affect read quality and the resulting allele calls in one strand direction. In contrast to our initial findings, of the three SNPs showing reads of one direction affected by context sequence features, only rs6875659 had a mean strand bias value outside the 25-75% range considered necessary for reliable genotyping [13].

3.5.2. SNPs with discordant genotypes and exclusion of component rs2080161

Four SNPs showed discordant genotypes in control DNA samples: rs9934011, rs9908046, rs6875659 and rs2080161. SNP rs595961 was discordant in several population samples. All discordances resulted from misalignments due to homopolymeric tracts in the context sequence. Such misalignments occurred only once for rs9934011 and rs9908046 (different control DNAs), likely a random effect. Genotypes in rs2080161, rs595961 and rs6875659 revealed discordances in different analyses of the same control sample or in several different population samples, indicating a systematic error. As discussed in section 3.2.4, allele calls for rs2080161, rs595961 and rs6875659 were all biased in one strand direction due to homopolymeric tracts. Typical IGV results for these SNPs are shown in Supplementary File S3. For example, rs595961 has two poly-C tracts of 5 and 4 consecutive Cs within 25 nucleotides of the SNP site, causing a proportion of reverse direction reads to be unreliable (Supplementary File S3, SNP 3). The same applies to rs6875659, with forward direction reads affected by poly-C and poly-A tracts (Supplementary File S3, SNP 2). Genotypes of both SNPs could be manually corrected by deduction of allele calls from just one read direction whereas rs2080161 was surrounded by several long (>5 nt) poly-T tracts in both directions (Supplementary File S3, SNP 1), which made manual correction impossible and led to the exclusion of this SNP.

3.5.3. SNPs with no-call genotypes

Ten SNPs gave higher than average no-calls (see Supplementary Table S1, row 9-18, column K, T and U) in control DNAs and population samples (>1 in 551 samples). The majority of no-calls were due to low coverage (<5-30x) or low quality variant

calls in certain samples, which is in accordance with minimum sequence coverage thresholds established for reliable SNP genotyping in other MPS sequencing studies [13,46-49].

Overall, extensive manual revision of raw sequencing output (BAM files) and Genotyper output (vcf files) as well as previously established thresholds of key sequence quality parameters [13] ensured all genotypes used for population analysis were reliable. Nevertheless, some of the SNPs that required manual checks are best replaced by loci having less problematic context sequence in future revisions of the Global AIM SNP panel. In addition, a high priority will be placed on the improvement and development of software analysis tools for MPS data to address the problem of homopolymeric tracts, since this context sequence characteristic creates one of the major challenges for MPS Indel detection and genotyping [48-50].

3.6. Population data and analyses

Summary allele frequencies for 127 Global AIM-SNPs estimated from 14 study populations are listed in Supplementary Table S5. Fig. 4 shows the analysis of these 14 study populations plus 1000 Genomes test populations extended to the additional data released after the first Global AIM-SNPs publication [17]. The average number of pairwise genotype differences between- (Fig. 4A, green cells) and within-populations (Fig. 4A, orange cells) plus pairwise F_{ST} values (Fig. 4A, blue cells) are outlined in Plot A. Using population ordering matched to Fig. 4A, principal component analyses (PCA) and optimum *K*=5 *STRUCTURE* cluster plots are shown in Fig. 4B and Fig. 4C, respectively.

Average pairwise genotype differences and F_{ST} distributions of the 32 test and study populations match well with the patterns that can be expected from their geographic distribution and admixture levels. Notable examples include: i. Somalis (SO) showing reduced differentiation with EUR populations compared to other AFR populations, with ASW and ACB also affected by degrees of EUR admixture; ii. Greenlanders (GL) least differentiated from admixed AMR populations, revealing a degree of shared AMR and EUR co-ancestry; iii. Fijian and East Timorese (FJ/ET) least differentiated from OCE; and iv. AMR admixed American region populations

showing the highest levels of within-population variation, with similar patterns seen in Indians, Afghanis (IN/AF) and Greenlanders (GL).

The *STRUCTURE* plot of Fig. 4C indicates cluster membership patterns similarly match well with each population's geographic distribution and likely admixture levels in almost all cases. Results shown in Fig. 4C are for the optimum *K*=5 inferred clusters using five reference populations. This is the chosen analysis option excluding 1000 Genomes South Asian data, as the differentiation of Europeans and South Asians was not a guiding factor in the original Global AIM-SNP selection. Calculating the cumulative PSD values for five vs. six comparisons reflects this marker selection, as South Asians only reach a cumulative PSD value of ~3 compared to ~11-15 for the other group comparisons (Supplementary Fig. S1A). Therefore, to provide adequate differentiation of Eurasian population groups, additional South Asian informative AIM-SNPs (such as those in Eurasiaplex [20]) will be required to re-balance the cumulative PSD values and ensure unbiased analysis of this sixth group. Nevertheless, an analysis of South Asian reference and test populations (Supplementary Fig. S1A) plus selected study populations was also performed and results are shown in Supplementary Fig. S2. Despite a much-reduced South Asian divergence in Global AIM-SNPs, results indicate an optimum *K*=6 was observed for six reference populations (Supplementary Fig. S1B). In the *K*=5 *STRUCTURE* plot of Fig. 4C we highlight the cluster patterns in Somalis and Greenlanders. First, Somalia positioned on the eastern edge of Africa has been subject to prolonged admixture with South Asian and Middle East populations. Somali samples show almost equal AFR and EUR cluster proportions corresponding to the expected co-ancestry patterns, treating Eurasia as one population group that includes all three sub-groups. However, when South Asian reference and test populations are used to analyze Somalis (Supplementary Fig. S2) European variation still forms the second cluster in the majority of samples. These *K*=6 results underline the lack of divergence between South Asians and Europeans, highlighting the need to reconfigure or supplement the Global AIM-SNPs to properly address this extra differentiation. Second, Greenlander samples show the most complex patterns of multiple cluster membership. This is likely to reflect their particular origin from Siberia and NE Asia migrations [51], which is different to the other American region populations analyzed, while they have also undergone recent additional European admixture.

For simplicity, Fig. 4B shows four PCA plots only indicating positions of reference (plots B1 and B2) and 1000 Genomes test populations (unadmixed B3; admixed B4). All PCA clusters are distributed in positions that correspond with the results of pairwise genotype difference and *STRUCTURE* analyses. The study populations have been individually plotted with one PCA per population in Supplementary Fig. S3 and with these analyses we have included *STRUCTURE* and ranked cross-validation likelihood ratio plots from *Snipper*-based Bayes analysis. Although likelihood ratios are generally uninformative for individuals with co-ancestry, i.e. they simply indicate an ancestry inference from the highest likelihood even when this is relatively low, ranking all the likelihoods obtained in a population sample in a $\log_{10}$LR plot can be instructive. In the study populations most likelihoods are well above the balanced odds line of LR=1 and only two GL samples fall below this line to be assigned as European, not Native American. In this and all other cases of the lowest likelihoods, the corresponding PCA points and *STRUCTURE* membership proportions are clearly discernible and indicate above average co-ancestry in these individuals compared to the other samples. Ranked $\log_{10}$LR plots are also shown for the six group comparisons in Supplementary Fig. S2, indicating IN, AF and IQ study populations are inferred to be South Asian and only five samples with higher likelihoods to be European (but all with low values). Therefore, ranked $\log_{10}$LR plots are a useful way to compare an individual's Bayes likelihoods with other samples used to make the ancestry analysis. Having incorporated Bayes and PCA analysis into a single portal in *Snipper*, we will further adapt the results output to indicate the position of the unknown SNP profile compared to the training set data in ranked plots to aid interpretation of samples with low likelihoods.

4. Concluding remarks

The incorporation of 125 of 128 original Global AIM-SNPs plus three substitute loci into an optimized PCR multiplex for the Ion PGM™ represents a successful porting of carefully selected markers into MPS analysis. This suggests new forensic SNP selections for extended ancestry analyses (likely necessary for the differentiation of

Middle East and South Asian populations from Europeans) or chosen to build novel forensic phenotyping tests, can be put into the target amplification step of MPS with a good chance of success. Most of the predictive SNPs forming forensic phenotyping tests are unlikely to be affected by low complexity flanking sequence, which is not found in coding regions. The fact that one AIM-SNP rs2080161 gave low quality sequence reads suggests context sequence scrutiny can be further improved during the Ion PGM™ custom assay design process. Overall, the loss of one AIM-SNP and replacement of another three had very little effect on the well balanced cumulative PSD values the Global AIM-SNPs panel sought to provide and this is underlined by results of analyses of several populations with complex admixture backgrounds.

Although all population analysis approaches used were able to differentiate South Asians from Europeans in nearly all cases, the much reduced cumulative PSD of this population group indicates that assessments of admixture involving South Asian co-ancestry, such as those made for Somalis, would not be free from estimation bias that may under-estimate the South Asian contribution compared to other population groups. Therefore, a new compilation of additional AIM-SNPs able to differentiate Eurasian sub-groups is a worthwhile next step to address the operational needs of forensic laboratories with a significant proportion of South Asian and Middle East populations in their region's demography. Furthermore, replacement of those few SNPs we found to require manual checks of genotyping accuracy will necessitate a careful adjustment of the panel's composition (and evaluation of new replacement SNPs). Some replacements could be relatively straightforward, for example rs2080162 has identical allele frequencies to the excluded rs2080161, but none of the flanking region polymeric nucleotide tracts that plague this SNP's sequence alignment.

Our studies took care to fully assess a set of mixtures, albeit from a single set of two-donor combinations. This was because ancestry-informative markers can potentially add useful information to the challenging task of de-convoluting mixed DNA patterns by allowing the inference of the ancestry of components in simple mixtures. MPS analysis provides a possibility to detect mixed DNA as the allele read frequency ratios give a reliable indicator of the genotype combinations in the sample. The results of our tests of mixed DNA with contrasted contributor ancestries are summarized in Fig.

2 and show a clear shift in homozygote ARFs away from each SNP's 0/1 baseline, as well as disrupted heterozygote balance, indicating the likely ancestry of the minor component in both 1:9 and 1:3 mixture ratios.

Acknowledgments

This work was funded by the European Union Seventh Framework Program (FP7/2007–2013) under grant agreement no. 285487 (EUROFORGEN-NoE) and the Austrian Science Fund (FWF) [P22880-B12]. CS is supported by funding awarded by the Portuguese Foundation for Science and Technology (FCT) and co-financed by the European Social Fund (Human Potential Thematic Operational Program SFRH/BD/75627/2010). MdlP is supported by funding awarded by the Consellería de Cultura, Educación e Ordenación Universitaria of the Xunta de Galicia as part of the Plan Galego de Investigación, Innovación e Crecemento 2011-2015 (Plan I2C).

References

[1] D.J. Allocco, Q. Song, G.H. Gibbons, M.F. Ramoni, I.S. Kohane, Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms, BMC Genomics 8 (2007) 68.

[2] M.A. Enoch, P.H. Shen, K. Xu, C. Hodgkinson, D. Goldman, Using ancestry-informative markers to define populations and detect population stratification, J. Psychopharmacol. 20 (2006) 19-26.

[3] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, R.M. Myers, Worldwide human relationships inferred from genome-wide patterns of variation, Science 319 (2008) 1100-1104.

[4] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, M.W. Feldman, Genetic structure of human populations, Science 298 (2002) 2381-2385.

[5] A. Manica, F. Prugnolle, F. Balloux, Geography is a better determinant of human genetic differentiation than ethnicity, Hum. Genet. 118 (2005) 366-371.

[6] N.A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J.K. Pritchard, M.W. Feldman, Clines, clusters, and the effect of study design on the inference of human population structure, PLoS Genet. 1 (2005) e70.

[7] D. Serre, S. Paabo, Evidence for gradients of human genetic diversity within and among continents, Genome Res. 14 (2004) 1679-1685.

[8] C. Phillips, Forensic genetic analysis of bio-geographical ancestry. Forensic Sci. Int. Genet. 18 (2015) 49-65.

[9] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet. 73 (2003) 1402-1422.

[10] M.D. Shriver, M.W. Smith, L. Jin, A. Marcini, J.M. Akey, R. Deka, R.E. Ferrell, Ethnic-affiliation estimation by use of population-specific DNA markers, Am. J. Hum. Genet. 60 (1997) 957-964.

[11] C. Børsting, S.L. Fordyce, J. Olofsson, H. Smidt Mogensen, N. Morling, Evaluation of the Ion Torrent™ HID SNP 169-plex: A SNP typing assay developed for human identification by second generation sequencing, Forensic Sci. Int. Genet. 12 (2014) 144–154.

[12] R. Daniel, C. Santos, C. Phillips, M. Fondevila, R.A. van Oorschot, Á. Carracedo, M.V. Lareu, D. McNevin, A SNaPshot of next generation sequencing, Forensic Sci. Int. Genet. 14 (2014) 50–60.

[13] M. Eduardoff, C. Santos, M. de la Puente, T.E. Gross, M. Fondevila, C. Strobl, B. Sobrino, D. Ballard, P.M. Schneider, Á. Carracedo, M.V. Lareu, W. Parson, C. Phillips, Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™, Forensic Sci. Int. Genet. 17 (2015) 110–121.

[14] S.L. Fordyce, H.S. Mogensen, C. Borsting, R.E. Lagace, C.W. Chang, N. Rajagopalan, N. Morling, Second-generation sequencing of forensic STRs using the

Ion Torrent HID STR 10-plex and the Ion PGM, Forensic Sci. Int. Genet. 14 (2015) 132-140.

[15] S.B. Seo, J.L. King, D.H. Warshauer, C.P. Davis, J. Ge, B. Budowle, Single base polymorphism typing with massively parallel sequencing for human identification, Int. J. Legal Med. 127 (2013) 1079–1086.

[16] C. Børsting, N. Morling, Next generation sequencing and its applications in forensic genetics, Forensic Sci. Int. Genet. 18 (2015) 78-89.

[17] C. Phillips, W. Parson, B. Lundsberg, C. Santos, A. Freire-Aradas, M. Torres, M. Eduardoff, C. Børsting, P. Johansen, M. Fondevila, N. Morling, P. Schneider; the EUROFORGEN-NoE Consortium; Á. Carracedo, M.V. Lareu, Building a forensic ancestry panel from the ground up: the EUROFORGEN Global AIM-SNP set, Forensic Sci. Int. Genet. 11 (2014) 13–25.

[18] B. Merriman, Ion Torrent R&D Team, J.M. Rothberg, Progress in Ion Torrent 650 semiconductor chip based sequencing, Electrophoresis 33 (2012) 3397–3417.

[19] M. Kayser, Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int. Genet. 18 (2015) 33–48.

[20] C. Phillips, A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M.D. Perez Carceles, Á. Carracedo, P.M. Schneider, M.V. Lareu, Eurasiaplex: A forensic SNP assay for differentiating European and South Asian ancestries, Forensic Sci. Int. Genet. 7 (2013) 359–366.

[21] C. Santos, C. Phillips, M. Fondevila, R. Daniel, R.A.H. van Oorschot, E.G. Burchard, M.S. Schanfield, L. Souto, J. Uacyisrael, M. Via, Á. Carracedo, M.V. Lareu, Pacifiplex: an ancestry-informative SNP panel centred on Australia and the Pacific region, Forensic Sci. Int. Genet. 20 (2016) 50–60.

[22] The 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes, Nature 491 (2012) 56-65.

[23] R. Drmanac, A.B. Sparks, M.J. Callow, A.L. Halpern, N.L. Burns, B.G. Kermani, P. Carnevali, I. Nazarenko, G.B. Nilsen, G. Yeung, et al., Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays, Science 327 (2009) 78-81.

[24] 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature 526 (2015) 68-74.

[25] J. Amigo, A. Salas, C. Phillips, Á. Carracedo, SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access, BMC Bioinformatics 9 (2008) 428.

[26] A. Onogi, M. Nurimoto, M. Morita, Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods, BMC bioinformatics 12 (2011) 263.

[27] B. Egyed, A. Brandstätter, J.A. Irwin, Z. Padar, T.J. Parsons, W. Parson, Mitochondrial control region sequence variations in the Hungarian population:

analysis of populations samples from Hungary and from Transylvania (Romania), Forensic Sci. Int. Genet. 1 (2007) 158–162.
[28] Thermo Fisher Scientific, Life Technologies: Ion AmpliSeq™ library preparation user guide. April 2014.
[29] Thermo Fisher Scientific, Life Technologies: Ion OneTouch™ 200 Template Kit v2 user guide. 2014.
[30] Thermo Fisher Scientific, Life Technologies: Ion PGM™ 200 Sequencing Kit user guide. 2014.
[31] Thermo Fisher Scientific, Life Technologies: Torrent Suite™ software 4.2 user guide. January 2015.
[32] R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Available from: http://www.R-project.org.
[33] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative Genomics Viewer, Nat. Biotechnol. 29 (2011) 24–26.
[34] H. Thorvaldsdóttir, J.T. Robinson, J.P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, Brief. Bioinform. 14 (2012) 178-192.
[35] R. Nassir, R. Kosoy, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, P.K. Gregersen, J.W. Belmont, F.M. De La Vega, M.F. Seldin, An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels, BMC Genetics 10 (2009) 39.
[36] http://mathgene.usc.es/snipper/analysispopfile2_new.html
[37] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945-959.
[38] L. Porras-Hurtado, Y. Ruiz, C. Santos, C. Phillips, Á. Carracedo, M.V. Lareu, An overview of STRUCTURE: applications, parameter settings, and supporting software, Front Genet 4 (2013) 98.
[39] D.A Earl, B.M. vonHoldt, STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method, Conservation Genet. Resour. 4 (2012) 359-361.
[40] G. Evanno, S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, Mol. Ecol. 14 (2005) 2611-2620.
[41] M. Jakobsson, N.A. Rosenberg, CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure, Bioinformatics 23 (2007) 1801-1806.
[42] N.A. Rosenberg, Distruct: a program for the graphical display of population structure, Molecular Ecology Notes 4 (2004) 137-138.
[43] L. Excoffier, H.E. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (2010) 564-567.

[44] http://browser.1000genomes.org/index.html, data for rs6875659 [accessed July 2015].
[45] http://browser.1000genomes.org/index.html, data for rs146348214 [accessed July 2015].
[46] D. Sims, I. Sudbery, N.E. Ilott, A. Heger, C.P. Ponting, Sequencing depth and coverage: key considerations in genomic analyses, Nat. Rev. Genet. 15 (2014) 121-132.
[47] R. Nielsen, J.S. Paul, A. Albrechtsen, Y.S. Song, Genotype and SNP calling from next-generation sequencing data, Nat. Rev. Genet. 12 (2011) 443-451.
[48] M.A. Quail, M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, P. Harold, H.P. Swerdlow, Y. Gu, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC Genomics 24 (2012) 341-353.
[49] D.R. Bentley, S. Balasubramanian, H.P. Swerdlow, G.P. Smith, J. Milton, C.G. Brown, K.P. Hall, D.J. Evers, C.L. Barnes, H.R. Bignell, et al., Accurate whole human genome sequencing using reversible terminator chemistry, Nature 456 (2008) 53–59.
[50] L.M. Bragg, G. Stone, M.K. Butler, P. Hugenholtz, G.W. Tyson, Shining a light on dark sequencing: Characterising errors in Ion Torrent PGM data, PLoS Comput. Biol. 9 (2013) e1003031.
[51] V. Colonna, L. Pagani, Y. Xue, C. Tyler-Smith, A world in a grain of sand: human history from genetic data, Genome Biol. 12 (2011) 234.

Figure legends

Fig. 1 Percentage of heterozygous (dark grey), homozygous (medium grey) and no-call (light grey) loci resulting from Genotyper calls for the 127 Global AIM-SNPs included in the EUROFORGEN Global AIMs panel. From left to right: (A) SS1 (single-source DNA sample 1 – European Coriell control sample NA07000 used as the first component of the mixtures), SS2 (single-source DNA sample 2 - African Coriell control sample NA18498 used as the second component of the mixtures) and Exp MIXT (theoretically calculated mixture using the single-source DNA genotypes); (B and C) different mixtures ratios (1:9, 1:3; 1:1, 3:1 and 9:1) analyzed in two different sequencing runs using Genotyper's default parameter settings and *minimum_allele_frequency* set to 0.02, respectively.

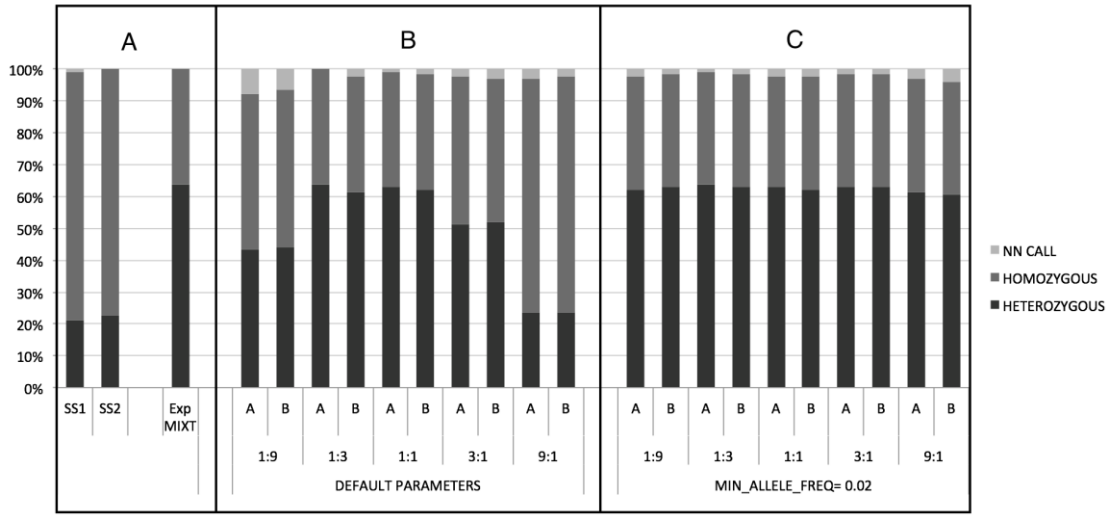


Fig. 2 Representation of the percentage of reads of a European sample's allele coverage divided by total coverage. From top to bottom and right to left: SS1 (European single-source DNA sample), mixtures from ratios 9:1, 3:1, 1:1,1:3 and 1:9 (the two runs are represented with different shades of grey) and SS2 (African single-source DNA sample).

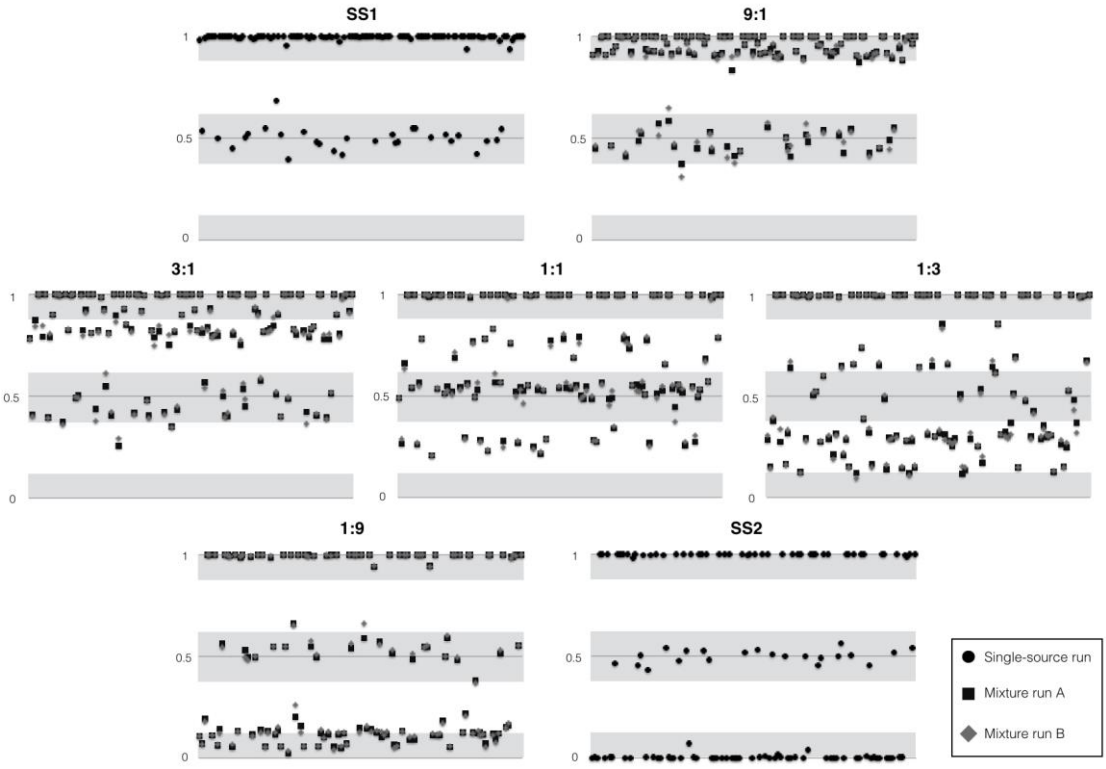
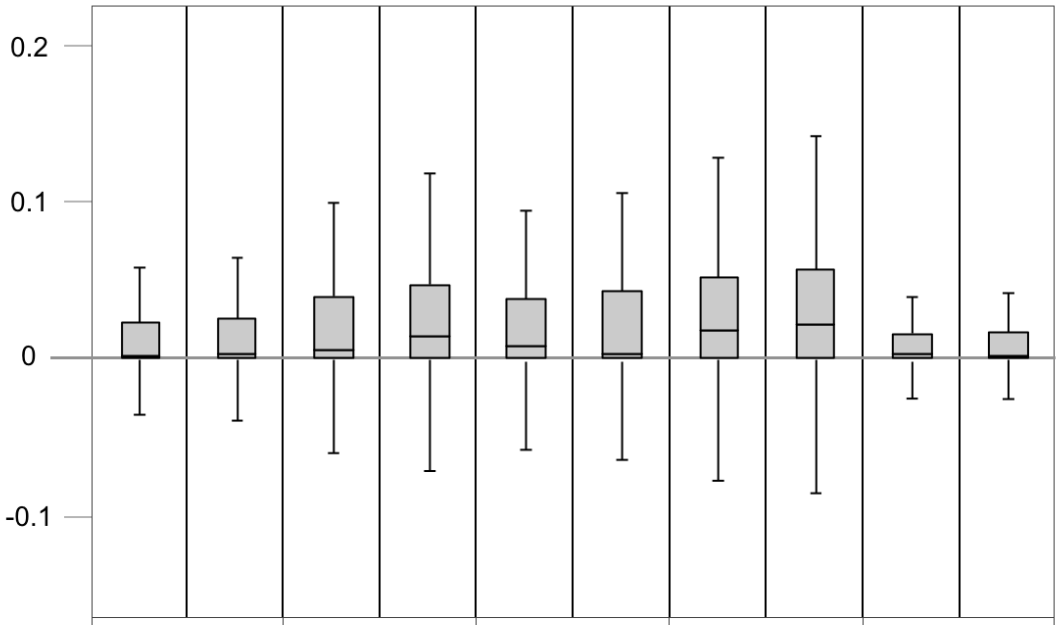


Fig. 3 Box plots representing observed minus expected allele read frequencies for the different ratios of mixtures, from 1:9 to 9:1, in both runs. In the table below, R^2 values for a lineal regression model, mean deviation values and standard deviations are also listed for each replicate.



	A	B	A	B	A	B	A	B	A	B
mean	0.012	0.013	0.022	0.023	0.020	0.020	0.025	0.027	0.005	0.006
sd	0.029	0.031	0.043	0.042	0.041	0.043	0.041	0.046	0.026	0.030
R^2	0.998	0.998	0.994	0.994	0.989	0.988	0.982	0.978	0.992	0.989

Fig. 4 Ancestry analyses of test and study populations compared with five reference populations. A) F_{ST} genetic distances and pairwise genotype differences within and between populations. Reference populations are indicated by a red square. B) PCA plot of the first two components (PC1 vs. PC2, plot 1) as well as the second and third components (PC2 vs. PC3, plot 2). Plot 3 shows unadmixed test populations and plot 4 admixed test populations for the first two components. Test individual coordinates were calculated from the principal components of reference samples. C) STRUCTURE analysis results. Optimum cluster number was *K*=5. Admixed test and study population individuals are ordered by decreasing value of major ancestry component. *Admixed test populations.

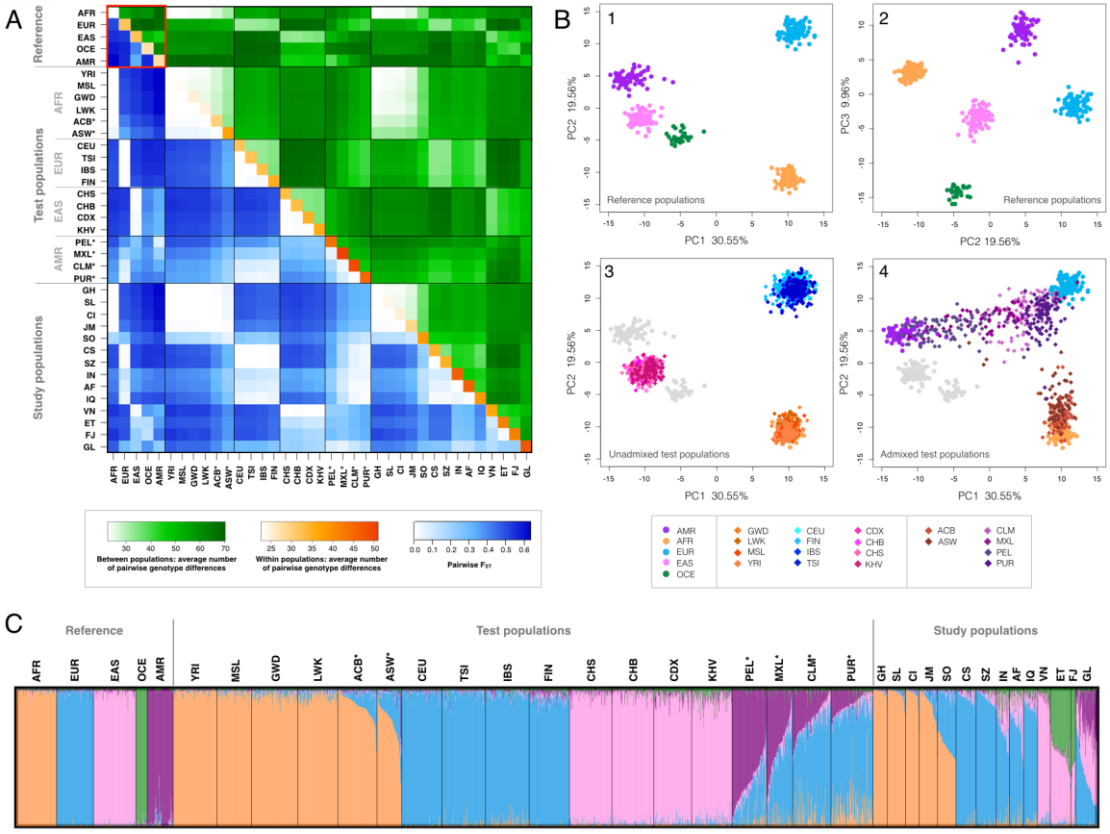


Table 1. Details of populations analyzed. Pop: population code; Group: population group; N: number of individuals. AFR: African; EUR European; EAS: East Asian; OCE: Oceanian; AMR: Native American; SAS: South Asian.

Set	No.	Pop	Group	*N*	Data Source	Description
Reference	1	**ESN**	AFR	99	1000 Genomes	Esan in Nigeria
	2	**GBR**	EUR	91	1000 Genomes	British in England and Scotland
	3	**JPT**	EAS	104	1000 Genomes	Japanese in Tokyo, Japan
	4	**OCE**	OCE	28	HGDP-CEPH	17 Papuan from New Guinea and 11 Melanesian from Bougainvillea
	5	**AMR**	AMR	64	HGDP-CEPH	14 Karitiana and 8 Surui from Brazil, 21 Maya and 14 Pima from Mexico, and 7 Piapoco from Colombia
	6	**GIH**	SAS	103	1000 Genomes	Gujarati Indian from Houston, Texas
Test	7	**YRI**	AFR	108	1000 Genomes	Yoruba in Ibadan, Nigeria
	8	**MSL**	AFR	85	1000 Genomes	Mende in Sierra Leone
	9	**GWD**	AFR	113	1000 Genomes	Gambian in Western Divisions in the Gambia
	10	**LWK**	AFR	99	1000 Genomes	Luhya in Webuye, Kenya
	11	**CEU**	EUR	99	1000 Genomes	Utah Residents with North and Western European ancestry
	12	**TSI**	EUR	107	1000 Genomes	Toscani in Italia
	13	**IBS**	EUR	107	1000 Genomes	Iberian Population in Spain
	14	**FIN**	EUR	99	1000 Genomes	Finnish in Finland
	15	**CHS**	EAS	105	1000 Genomes	Southern Han Chinese
	16	**CHB**	EAS	103	1000 Genomes	Han Chinese in Beijing, China
	17	**CDX**	EAS	93	1000 Genomes	Chinese Dai in Xishuangbanna, China
	18	**KHV**	EAS	99	1000 Genomes	Kinh in Ho Chi Minh City, Vietnam
	19	**PJL**	SAS	96	1000 Genomes	Punjabi from Lahore, Pakistan
	20	**BEB**	SAS	86	1000 Genomes	Bengali from Bangladesh
	21	**STU**	SAS	102	1000 Genomes	Sri Lankan Tamil from the UK
	22	**ITU**	SAS	102	1000 Genomes	Indian Telugu from the UK
	23	**ACB**	Admixed	96	1000 Genomes	African Caribbeans in Barbados
	24	**ASW**	Admixed	61	1000 Genomes	Americans of African Ancestry in SW USA
	25	**PEL**	Admixed	85	1000 Genomes	Peruvians from Lima, Peru
	26	**MXL**	Admixed	64	1000 Genomes	Individuals with Mexican Ancestry from Los Angeles USA
	27	**CLM**	Admixed	94	1000 Genomes	Colombians from Medellin, Colombia
	28	**PUR**	Admixed	104	1000 Genomes	Puerto Ricans from Puerto Rico
Study	29	**GH**	AFR	35	Present study	Ghana
	30	**SL**	AFR	45	Present study	Sierra Leone
	31	**CI**	AFR	33	Present study	Ivory Coast
	32	**JM**	Admixed	45	Present study	Jamaica
	33	**SO**	AFR	46	Present study	Somalia

	34	**CS**	EUR	49	Present study	Csango from Hungary
	35	**SZ**	EUR	50	Present study	Szeklers from Hungary
	36	**IN**	SAS	32	Present study	India
	37	**AF**	SAS	35	Present study	Afghanistan
	38	**IQ**	Middle East	34	Present study	Kurdish from Iraq
	39	**VN**	EAS	32	Present study	Vietnam
	40	**ET**	OCE	50	Present study	East Timor
	41	**FJ**	OCE	12	Present study	Fiji
	42	**GL**	AMR	53	Present study	Greenlanders

Table 2 Concordance details for comparisons of Ion PGM™ genotype calls from five laboratories and online data for Coriell control DNA samples.

SNP ID	Coriell control DNA sample No.	No. of analyses with discordance	Discordant genotype	Concordant lab genotypes	Complete Genomics genotype	1000 Genomes-Phase III genotype	Comments on discordance
rs6875659	CTR_NA18498	4/5	AG	AA	AA	AA	See section 3.5.2
rs2080161	CTR_NA11200	3/5	AC	CC	-	-	See section 3.5.2.
rs9934011	CTR_NA11200_lab3	1/5	CT	CC	-	-	See section 3.5.2.
rs9908046	CTR_NA10540_lab3	1/5	CT	TT	-	-	See section 3.5.2.

Table 3 Average sequence coverage, number of no-call genotypes, allele drop-in, allele drop-out and locus drop-out for samples typed with DNA input amounts of 50 pg, 25 pg and 10 pg, plus degraded DNA sample Bone 2.

	50pg, 25 cycles	50pg, 25+5 cycles	25pg, 25 cycles	25pg, 25+5 cycles	10pg, 25+5 cycles	Degraded DNA sample Bone 2
Average Coverage	265	183	267	211	48	430
No-call genotype	3	3	4	3	9	4
Allele drop-in	1	0	2	1	1	0
Allele drop-out	1	2	5	10	7	0
Locus drop-out	0	0	1	15	49	0
Total	5	5	12	29	66	4